APPENDIX B. INFORMATION QUALITY SOFTWARE TOOLS

Information quality tools provide automation and management support for solving information quality problems (source: *Improving Data Warehouse and Business Information Quality*, Chapter 10). Effective use of information quality tools requires

- Understanding the problem you are solving,
- Understanding the kinds of technologies available and their general functionality,
- Understanding the capabilities of the tools,
- Understanding any limitations of the tools,
- Selecting the right tools based on your requirements,
- Using the tools properly.

Sections B.1-B.5 below discuss five categories of information tools for information quality improvement and data correction that may be applied within individual Program Areas at HUD to support the four-stage process for information quality improvement discussed in this Handbook. It is recommended that each Program Area choose its own tools to support specific program business needs. It is always recommended that a Program Area select only one tool to accomplish a single category of information quality improvement.

The caveat for all automated correction tools is that some varying percentage of the data will need to be corrected and verified manually by looking at hard copy "official" documents or by comparing it to the real world object or event. Also, automated tools cannot ensure "correctness" or "accuracy."

B.1 INFORMATION QUALITY ANALYSIS TOOLS

Automated tools may be used to conduct audits of data against a formal set of business rules to discover inconsistencies within those rules. Reports can be generated that depict the number and type of errors found. Quality analysis and audit tools measure the state of conformance of a database or process to the defined business rule.

B.2 BUSINESS RULE DISCOVERY TOOLS

Business rule discovery tools may be used to analyze legacy system data files and databases in order to identify data relationships affecting the data. This analysis may identify quantitative (formula-based) or qualitative (relationship-based) conditions affecting the data and its successful migration and transformation. The analysis may also uncover exceptions or errors in the conditions.

Business rule discovery tools use data mining or algorithms to analyze data to discover

- Domain value counts,
- Frequency distributions of data values,
- Patterns of data values in non-atomic data, such as unformatted names and addresses or textual data,
- Formulas or calculation algorithms,
- Relationships, such as duplicate data within or across files,
- Similarities of items, such as spelling,
- Correlation of data values in different fields,
- Patterns of behavior that may indicate possible fraud, intentional or unintentional.

It is important to remember that there may be performance problems when using these tools if the files are large or contain many fields. Performance problems may be minimized through random sampling or by making separate analysis runs against different sets of fields, grouped in ways that meaningful business rules are likely to emerge.

B.3 DATA REENGINEERING AND CORRECTION TOOLS

Data reengineering and correction tools may be used either to actually correct the data or to flag erroneous data for subsequent correction. These tools require varying degrees of in-house data knowledge and analysis to adequately use them. Data correction tools may be used to standardize data, identify data duplication, and transform data into a correct set of values. These tools are invaluable in automating the most tedious facets of data correction.

Data reengineering and correction tools may perform one or more of the following functions:

- Extracting data.
- Standardizing data.
- Matching and consolidating duplicate data.
- Reengineering data into architected data structures.
- Filling in missing data, based upon algorithms or data matching.
- Applying updated data, such as address corrections from change of address notifications.
- Transforming data values from one domain set to another.
- Transforming data from one data type to another.
- Calculating derived and summary data.
- Enhancing data, by matching and integrating data from external sources.
- Loading data into a target data architecture.

B.4 DEFECT PREVENTION TOOLS

Automated tools may also be used to prevent data errors at the source of entry. Application routines can be developed that test the data input. Generalized defect prevention products enable the definition of business rules and their invocation from any application system that may use the data. These tools enforce data integrity rules at the source of entry, thereby preventing problems before they occur.

Defect prevention tools provide the same kind of functions as data correction tools. The difference is that they provide for discovery of the errors and correction of them during the online data creation process, rather than in batch mode.

B.5 METADATA MANAGEMENT AND QUALITY TOOLS

Metadata management and quality tools provide automated management and quality control of data definition and information architecture development. The tools perform one or more of the following functions:

- Ensure conformance to data naming standards.
- Validate data name abbreviations.
- Ensure all required components of data definition are provided.
- Maintain metadata for control of data reengineering and correction processes.
- Evaluate data models for normalization.
- Evaluate database design for integrity, such as primary key to foreign key integrity, and performance optimization.

Metadata management and quality tools support the documentation of the specification of the information product. These tools cannot determine if data required for knowledge workers is missing, defined correctly, or even required in the first place. Information resource data (metadata) quality tools may audit or ensure that data names and abbreviations conform to standards, but they cannot assess whether the data standards are "good" standards that produce data names that are understandable to knowledge workers.

B.6 EVALUATING INFORMATION QUALITY TOOLS

Tool selection is second only to the business problem at hand in architecting a business solutions environment. Evaluate any software tool from the standpoint of how well it solves business problems and supports accomplishing the enterprise business objectives. Avoid "vendor pressure" to buy tools *before* you develop requirements.

Once you understand the business problems you are solving, determine what category of information quality function automation is required. For example, the fact that you are developing a data warehouse does not automatically mean that your problem is correcting data for the warehouse. The real problem may be data defects at the source, and the business problem to be solved is that the information producers do not know who uses the information they create. Therefore, a data defect prevention tool is required to solve the real business problem.

[THIS PAGE INTENTIONALLY LEFT BLANK]